PR #40349 完整报告

vllm-project/vllm

[Bugfix][CI] Fix `tests/distributed/test_torchrun_example_moe.py`

合并时间：2026-04-21 02:05

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40349

执行摘要

该 PR 修复了分布式 MoE 测试脚本中缺失的 max_model_len 和 max_num_seqs 配置参数，解决了因 LLM 初始化不完整导致的测试失败问题。这是一个简单的测试修复，仅影响特定测试用例，无生产代码风险。

功能与动机

根据 PR 描述，此变更旨在修复外部 issue（neuralmagic/vllm-ci-bugs#12）中报告的测试失败问题。测试脚本 test_torchrun_example_moe.py 在创建 LLM 实例时缺少必要的配置参数，导致测试无法正常执行。添加这两个参数是为了确保测试能正确初始化并验证分布式环境下 KV 缓存配置的一致性。

实现拆解

1. 变更入口：仅修改了 tests/distributed/test_torchrun_example_moe.py 文件，这是分布式 MoE 测试的入口脚本。
2. 核心修复：在 LLM 构造函数调用中增加了两个关键参数：
   - max_model_len=1024：控制模型支持的最大序列长度，直接影响 KV 缓存块数量的计算。
   - max_num_seqs=16：设置最大并发序列数，影响调度器的队列大小和内存预分配。
3. 测试逻辑保持：测试的其他逻辑（如数据并行提示分配、广播验证 KV 缓存一致性）完全不变，修复仅针对初始化配置。

关键代码片段展示了修复后的 LLM 初始化部分：

1. 无配套改动：没有涉及其他源码、配置文件或部署脚本的修改。

关键源码片段

tests/distributed/test_torchrun_example_moe.py

这是唯一被修改的文件，修复了分布式 MoE 测试的配置缺失问题。

```
llm = LLM(
    model="microsoft/Phi-mini-MoE-instruct",
    tensor_parallel_size=int(os.getenv("TP_SIZE", "1")),
    pipeline_parallel_size=int(os.getenv("PP_SIZE", "1")),
    enable_expert_parallel=int(os.getenv("ENABLE_EP", "0")) == 1,
    distributed_executor_backend="external_launcher",
```

```
    gpu_memory_utilization=random.uniform(0.7, 0.9),
    seed=0,
    max_model_len=1024, # 新增：设置最大模型长度，影响 KV 缓存块计算
    max_num_seqs=16, # 新增：设置最大并发序列数，影响调度队列大小
)
```

评论区精华

review 中没有实质性技术讨论。两个 bot 评论（claude[bot] 和 gemini-code-assist[bot]）均为自动化提示，未提供具体反馈。维护者 khluu 直接批准了 PR，表明变更被认可为简单且必要的修复。

风险与影响

风险分析：

- 低风险：仅修改测试脚本，不涉及生产代码，无回归风险。
- 添加的参数是 LLM 的标准配置项，值（1024 和 16）合理，不会引入兼容性问题。
- 修复后测试能更可靠地验证分布式一致性，有助于发现潜在问题。

影响分析：

- 对用户无直接影响，纯测试修复。
- 提升 CI 流水线稳定性，减少因测试配置缺失导致的失败。
- 影响范围仅限于特定分布式 MoE 测试用例。

关联脉络

从近期历史 PR 看，该修复属于测试维护范畴，与核心功能演进无直接关联。类似测试修复在仓库中常见（如 PR #40161 修复 CPU 资源探测、PR #39627 启用 XPU 测试），反映了团队对测试覆盖率和 CI 稳定性的持续关注。该 PR 针对 v0.20.0 里程碑，属于向后兼容的 bugfix。